WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Named Entity Recognition in Hindi-English Code-Mixed tweets

Kaberi Sangma[a], Shatasree Das[b], Dr. Amit Majumder[c*]

[a]Student, Department of CSE, JIS College Of Engineering, Kalyani
[b] Student, Department of CSE, JIS College Of Engineering, Kalyani
[c]Assistant Professor, Department of CSE, JIS College of Engineering, Kalyani

A B S T R A C T

Around 23.6 million Indian users are respectively active on Twitter, where users can tweet their points of view about many aspects. A maximum of these users is tended to know more than one language and these users can tweet in monolingual English or in their native language. Users are easily code-mixing of Hindi and English together or even trilingual to express their opinion on tweeter. Code-mixing is when a person uses more than one language while communicating. For this challenge, an automatic language identification tool becomes a very mandatory role for scanning the noisy content on tweeter or any other social media platforms. This project is on code-mix Extraction of Hindi and English. Our project presents a corpus for Named Entity Recognition (NER) in Hindi-English Code-Mixed along with extensive experiments on our machine learning model, where we have used a Decision Tree Algorithms, CRF, and SGD classifiers.

Keywords: Machine Learning, Named Entity Recognition, Natural Language Processing, SGD, CRF

1. INTRODUCTION

Code-mixing can usually see in bilingual and multilingual communities and the reason behind this is that they can easily express their ideas and thoughts in a specific language. Indian people speak many languages because of regional diversities, we can say that India is a land of many tongues. There are 23.6 million Indian users active on Twitter, where these users share their thoughts and ideas on Twitter and these users can tweet in monolingual English or in their native language. Users are easily code-mixing of Hindi and English together or even trilingual to express their opinion on tweeter or other social media platforms.

The following are some examples of code-mixing tweets.

Code-Mixing: "Achaa hua tumne hamare proposal accept kiya or else tumne toh pura paisa betting pe dalta."

Translation: "It is good that you have accepted ourproposal or else you would have spent all your money on betting."

Code-Mixing: "Zindagi usiko jeena ata hain who faces every obstacle of life."

Translation: "He knows how to live who faces every obstacle of life"

Code-mixing of tweets on Twitter can cause a serious issue to both the Online Social Networks (OSNs') basic text data mining algorithms along with researchers trying to study online discussion. hence most of the existing tools for analyzing Online Social Network (OSN) text content caters to monolingual data.

In this project our main contributions are:

1.Developing a token-level language recognition system for Hindi-English code-mixed or monolingual tweets.

2.Developing Named Entity Recognition (NER) for Hindi-English code-mixed tweets.

2. RELATED WORK

Identification of code-mixed Language content has been formerly explored [1], they addressed the problem of language identification on Bengali, Hindi, and English Facebook comments. In our project, we mainly focus on efforts to build Named Entity Recognition for Twitter or any other social network content, and for code-mixed languages corpora, we have used Named Entity Recognition. Tweets and any other Social Media text, have subtle variations from written and spoken text mainly tweets. Which are spelling variations, ad-hoc, slacker grammaticalstructure, and more. We can also see in detail the differences between traditional textual sources and tweets [3]. We can also see the extracting named entity and identifying its type using four languages namely English, Hindi, Malayalam, and Tamil [5]. or the challenges and problems in the Hindi English code-mixed text, Part-of-speech (POS) tag

annotated Hindi English code-mixed corpus have been performed [7]. We can also see that they have performed experiments on transliteration, language identification, normalization, and Part-of-speech (POS) tagging of the Dataset

3. METHODOLOGY

Before analyzing and processing social media text the pre-processing task is a necessary step. At first, on the given training data tokenization is executed. The tokens collected later during the tokenization process are converted into standard BIO format. This leads to BIO tag information for each word in the training data. Basically, the BIO (Beginning Inside Outside) tag. Let us know more about the BIO tag with the help of a simple example: "Mark Zuckerberg is Chief Executive Officer of Facebook". Generally, the entity 'Mark Zuckerberg' indicates PERSON and 'Facebook' indicates ORGANIZATION. Hence the word Mark Zuckerberg has two parts, it is tagged in BIO format as beginning and inside.

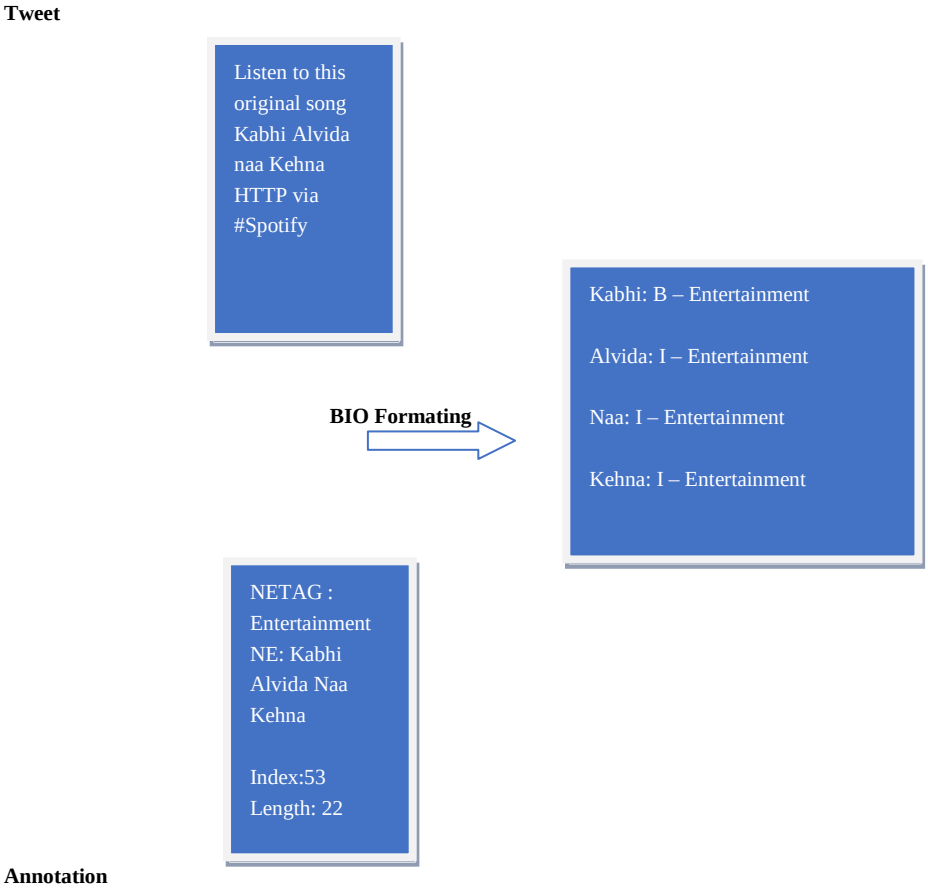
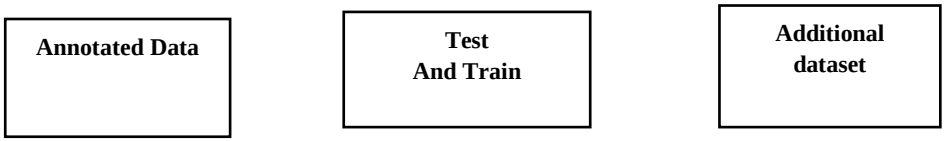


FIGURE.1. BIO-Formatting

An example of BIO-formatting is given in Figure.1.
An O tag is for the words which do not belong to any entity or chunk. We can see in the given example, that Mark is labeled as B-PERSON, Zuckerberg as IN-PERSON, and Facebook as ORGANIZATION



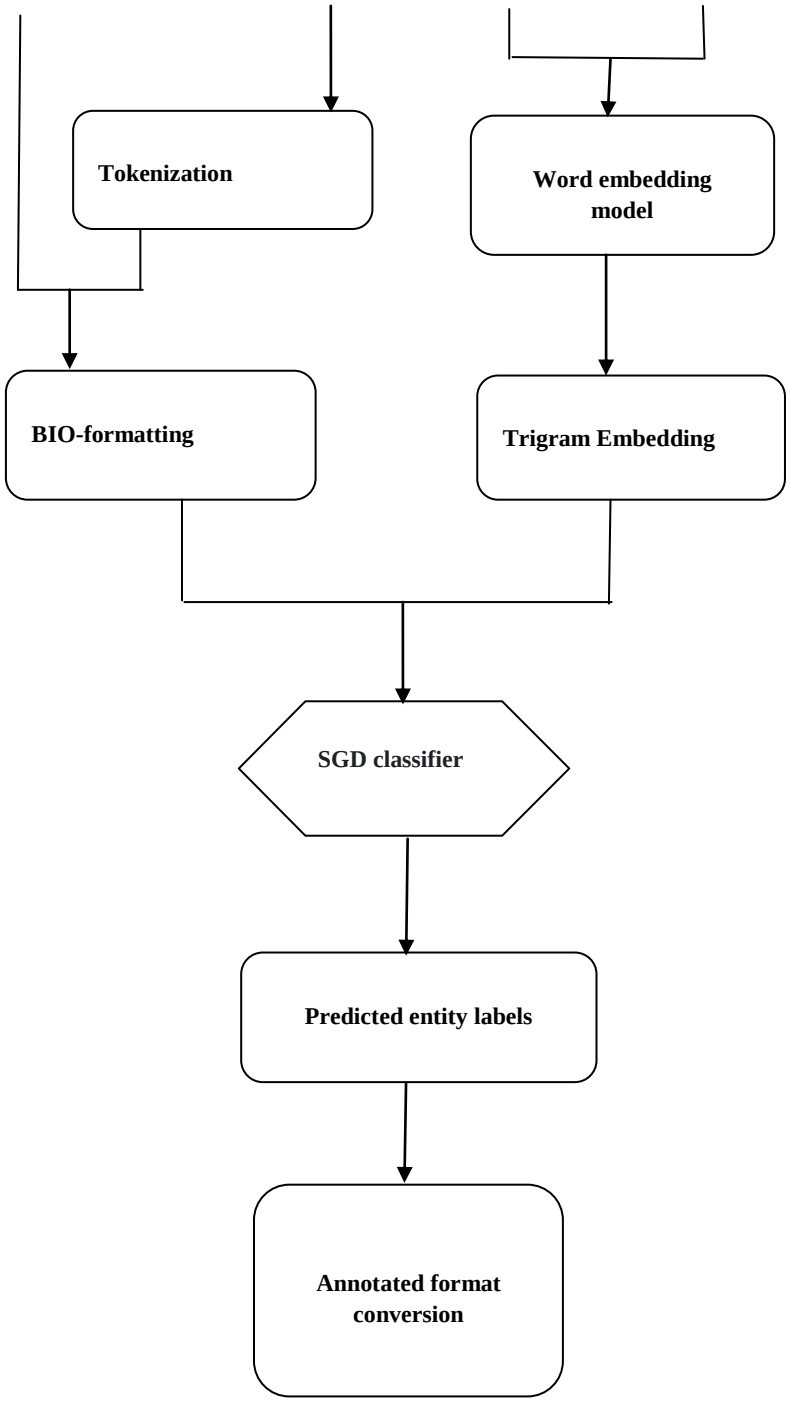


Figure.2.Word embedding models for Entity extraction system

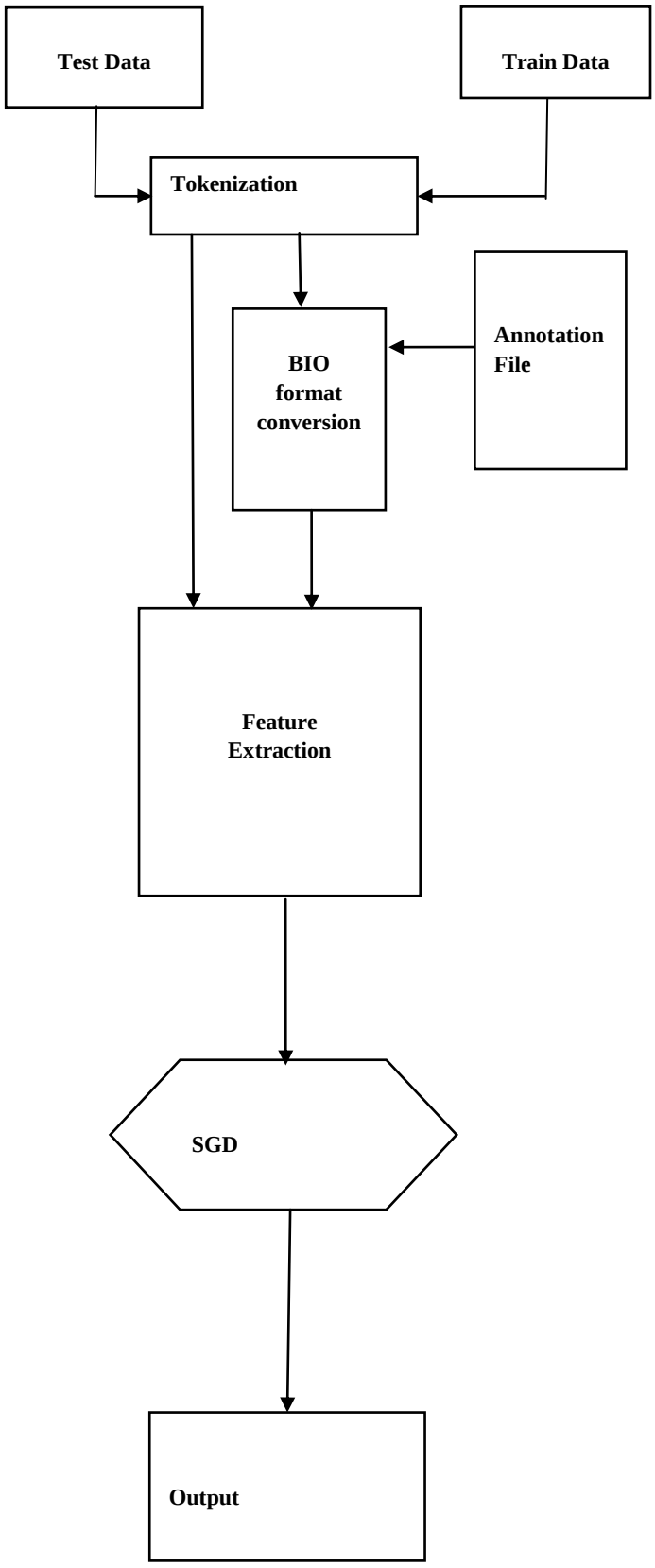


Figure.3. Stylometric features for Entity Extraction System.

4. RESULTS AND DISCUSSION

In this project, we haphazardly sampled 10,000 tweets of the code-mixed tweet from a collected dataset which is available on Kaggle. After that, using a language detection algorithm and we filtered tweets. The data which have been filtered also had tweets containing Roman tokens belonging to languages other than English and Hindi, such tweets were removed during the annotation process. In this dataset we can see every tweet is annotated at a token level for three named entity types i.e., Person, Location, and Organization, with the use of BIO format, BIO tagged sentences are given in Table 1.

Text	Label
In	O
Beirut	B-geo
,	O
a	O
string	O
of	O
officials	O
voiced	O
their	O
anger	O
,	O
while	O
at	O
the	O
United	B-org
Nations	I-org
summit	O
in	O
New	B-geo
York	I-geo
,	O
Prime	B-per
Minister	O
Fouad	B-per
Siniora	I-per
said	O
the	O

Table 1: BIO-tagged Sentences

Tag	Precision	Recall	F1- score
B-art	0.50	0.11	0.19
B-eve	0.62	0.24	0.35
B-geo	0.67	0.89	0.76
B-gpe	0.94	0.73	0.82
B-nat	0.43	0.16	0.24
B-org	0.63	0.51	0.57
B-per	0.81	0.56	0.66
B-tim	0.90	0.66	0.76
I-art	0.60	0.05	0.09
I-eve	0.75	0.09	0.16
I-geo	0.73	0.56	0.63
I-gpe	0.50	0.05	0.08
I-nat	0.25	0.10	0.14
I-org	0.73	0.52	0.61
I-per	0.65	0.73	0.69
I-tim	0.52	0.01	0.02
O	0.98	1.00	0.99
Avg/total	0.94	0.94	0.94

Table 2: Results using SGD classifier

Tag	Precision	Recall	F1-score
B-art	0.67	0.40	0.50
B-eve	0.00	0.00	0.00
B-geo	0.76	0.90	0.82
B-gpe	0.95	0.89	0.92
B-nat	1.00	1.00	1.00
B-org	0.71	0.53	0.60
B-per	0.85	0.89	0.87
B-tim	0.97	0.77	0.86
O	0.98	0.99	0.98
Avg/total	0.94	0.94	0.94

Table 3: Results using CRF classifier

5. COMPARISON OF EXPERIMENTAL RESULTS WITH EXISTING SYSTEM

In this project, our system can learn the particular Named Entity types such as Person, Location, and Organization from the form of the text with the use of BIO format as shown in Table 1, where we can see our system can finely understand because as we can see it is tagged the mass tokens precisely. We can see the results in Table 2 and Table 3. Where Table 2 is the Observation of the SGD classifier and Table 3 is the Observation of the CRF classifier where we can see our system determines the tag tokens which is the beginning of a Named Entity but the maximum time it is tagging as B-per that is the main problem. What our system needed to learn is to further common details over these particular characters. Besides B-per tags for other tags our system gives precise predictions

6. CONCLUSION AND FUTURE SCOPE

Our project presents a Named Entity Recognition tool mainly for Hindi and English code-mixed content. To develop our Named Entity Recognition model. Also, in this project to utilize the character-level differences in languages, we present a different semi-supervisedlanguage identifier and we authenticatethe execution of our Named Entity Recognition as opposed to off-the-shelf Named Entity Recognition for Twitter and detect that our model outperforms them.

Near future, we are planning to build other downstream Natural Language Processing tools, like Entity-specific Sentiment Analyzers which build the use of Named Entity Recognition for code-mixed data.

REFERENCES

[1] Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In Proceedings of the first workshop on computational approaches to code switching. pages 13– 23.

[2] Pius von Daniken and Mark Cieliebak. 2017. Trans- ¨ fer learning and sentence level features for named entity recognition on tweets. In Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 166–171.

[3] Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pages 1524– 1534.

[4] Aaron Jaech, George Mulcaire, Mari Ostendorf, and Noah A Smith. 2016. A neural model for language identification in code-switched tweets. In Proceedings of The Second Workshop on Computational Approaches to Code Switching. pages 60–64

[5] M. Anand Kumar, S. Shriya, and K. P. Soman. AMRITA-CEN@FIRE 2015: Extracting entities for social media texts in Indian languages. CEUR Workshop Proceedings, 1587:85–88, 2015.

[6] Deepak Gupta, Shubham Tripathi, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A hybrid approach for entity extraction in code-mixed social media data. MONEY, 25:66

[7] Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 974–979.

[8] Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In Proceedings of the Forum for Information Retrieval Evaluation. ACM, New York, NY, USA, FIRE '14, pages 48– 53. https://doi.org/10.1145/2824864.2824872.

[9] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. InProceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '05, pages 363–370. https://doi.org/10.3115/1219840.1219885.

[10] Kanika Gupta, Monojit Choudhury, and Kalika Bali. 2012. Mining hindi-english transliteration pairs from online hindi lyrics. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2012). pages 2459–2465.

[11] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. CoRR abs/1603.01360. http://arxiv.org/abs/1603.01360.

[12] Gyorgy Szarvas, Rich ¨ ard Farkas, and Andr ´ as Kocsor. ´ 2006. A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. In International Conference on Discovery Science, pages 267–278. Springer.

Authors Profile

Kaberi Sangma is currently doing her B.Tech from Computer Science and Engineering department, JIS College of Engineering, West Bengal, India. She hasinterest in Natural Language Processing, Machine Learning, and Information Extraction.

Shatasree Das is currently doing her B.Tech from Computer Science and Engineering department, JIS College of Engineering, West Bengal, India. She hasinterest in Machine Learning, Information Extraction, and Natural Language Processing.

Dr. Amit Majumder has received B.Tech degree in Computer Science and Engineering from Kalyani Govt. Engineering College, Kalyani, West Bengal, India, ME degree in Computer Science and Engineering from Jadavpur University, West Bengal, India and Ph.D. from Jadavpur University, West Bengal, India. Currently, he is working as Assistant Professor in CSE department at JIS College of Engineering, Kalyani, West Bengal, India. He has an interest in areas of Artificial Intelligence, Natural Language Processing, Machine Learning, Deep Learning, Computer Vision, and Information Extraction.